

BRL R 1369

BRL

AD

AD658674

REPORT NO. 1369

ON STEPWISE MULTIPLE LINEAR REGRESSION

by

Harold J. Breaux

August 1967

SEP 25 1967

Distribution of this document is unlimited.

U. S. ARMY MATERIEL COMMAND
BALLISTIC RESEARCH LABORATORIES
ABERDEEN PROVING GROUND, MARYLAND

CLEARINGHOUSE

53

Destroy this report when it is no longer needed.
Do not return it to the originator.

The findings in this report are not to be construed as
an official Department of the Army position, unless
so designated by other authorized documents.

BALLISTIC RESEARCH LABORATORIES

REPORT NO. 1369

AUGUST 1967

ON STEPWISE MULTIPLE LINEAR REGRESSION

Harold J. Breaux

Computing Laboratory

This report is based on a master's thesis presented to the University of Delaware, Department of Statistics and Computer Science, June, 1967.

Distribution of this document is unlimited.

RDT&E Project No. 1P014501A14B

ABERDEEN PROVING GROUND, MARYLAND

BALLISTIC RESEARCH LABORATORIES

REPORT NO. 1369

HJBreaux/bj
Aberdeen Proving Ground, Md.
August 1967

ON STEPWISE MULTIPLE LINEAR REGRESSION

ABSTRACT

Stepwise multiple linear regression has proved to be an extremely useful computational technique in data analysis problems. This procedure has been implemented in numerous computer programs and overcomes the acute problem that often exists with the classical computational methods of multiple linear regression. This problem manifests itself through the excessive computation time involved in obtaining solutions to the $2^N - 1$ sets of normal equations that arise when seeking an optimum linear combination of variables from the subsets of the N variables. The procedure takes advantage of recurrence relations existing between covariances of residuals, regression coefficients, and inverse elements of partitions of the covariance matrix. The application of these recurrence formulas is equivalent to the introduction or deletion of a variable into a linear approximating function which is being sought as the solution to a data analysis problem. This report contains derivations of the recurrence formulas, shows how they are implemented in a computer program and includes an improved algorithm which halves the storage requirements of previous algorithms. A computer program for the BRLESC computer which incorporates this procedure is described by the author and others in a previous

report, BRL Report No. 1330, July 1966. The present report is an amplification of the statistical theory and computational procedures presented in that report in addition to the exposition of the improved algorithm.

TABLE OF CONTENTS

| | Page |
|---|------|
| ABSTRACT. | 3 |
| I. INTRODUCTION. | 7 |
| II. MULTIPLE LINEAR REGRESSION. | 11 |
| III. COMPUTATIONAL CONSIDERATIONS IN MULTIPLE LINEAR REGRESSION. | 15 |
| IV. MATHEMATICAL BASIS OF THE STEPWISE REGRESSION | 17 |
| Derivation of Recurrence Formulas. | 20 |
| Elements of the Inverse Matrix | 23 |
| List of Recurrence Formulas. | 28 |
| Theorem on Stepwise Multiple Linear Regression | 29 |
| The Correlation Matrix | 32 |
| V. SELECTING THE KEY VARIABLE. | 34 |
| VI. IMPROVEMENT OF THE ALGORITHM. | 38 |
| VII. A COMPARISON OF FORWARD AND BACKWARD STEPWISE REGRESSION. | 42 |
| REFERENCES. | 46 |
| APPENDIX. | 49 |
| Numerical Example. | 49 |
| Recent Work In Europe. | 52 |
| DISTRIBUTION LIST | 55 |

I. INTRODUCTION

The computational technique for stepwise multiple linear regression described by M. A. Efroymson [5]* has proved to be extremely useful in data analysis problems. This procedure, with various modifications, has been implemented in numerous computer programs in government laboratories, universities, and industry and overcomes one of the major problems that often exists with the classical** computational methods of multiple linear regression. In problems where many variables are involved, one may have only intuitive suspicion regarding those variables which may be significant. In these instances, one of the classical approaches is to obtain the least-squares solution to the regression equation containing all the variables that are believed to be potentially significant and then attempt to eliminate insignificant variables by tests of significance. This procedure is of limited use when many variables are involved and usually runs into extreme computational difficulty. An alternative procedure is to examine the solutions of all the subset models that can

*Numbers in brackets denote references which may be found on page 46.

**The word "classical" here may be a misnomer in that the essential substance of the computational procedure was proposed as early as 1934 by Horst [12] and 1938 by Cochran [4]. The recent interest in the subject is of course due to the advent of modern high speed Computing machinery.

be formed from the collection of variables that are of interest and choose the one which seems to give the "best fit." This procedure, however, can be very costly in terms of computation time. If one has N independent variables and wishes to obtain all possible solutions to models containing 1, 2, ... and N variables one has to solve $2^N - 1$ sets of linear equations. For candidate models containing five variables this would require the solution of 31 sets of linear equations (a practical number) but for twenty variables this number jumps to 1,048,575. A means to circumvent this computational difficulty is provided by stepwise multiple regression. This procedure takes advantage of the fact that the Gauss-Jordan algorithm, when used to solve the normal equations with N variables, yields intermediate solutions to N regression problems containing 1, 2, ... and N variables. The power of the procedure lies in the fact that the variables are introduced into the regression in the order of their significance. At each stage the variable which is entered into the regression is the one which will yield the greatest reduction in the sum of squares of residuals. The power of the procedure is further enhanced by removing terms from regression at later stages that have become insignificant as a result of the inclusion of additional variables in the regression. The computations proceed until an equilibrium point is reached where no significant reduction in the sum of squares of residuals is to be gained by adding variables in the regression and where a significant increase in the sum of squares of residuals would arise if a variable were removed from regression. The procedure described above will be

referred to as forward stepwise regression. A modification of the method is to begin with all variables in regression and then remove insignificant variables, one by one. In a fashion similar to the forward regression, a variable which is removed from regression can subsequently reenter if it becomes significant at a later stage. This procedure will be referred to as backwards stepwise regression.

The optimum or ideal sub-model chosen from a candidate model can be defined as that model containing only variables which are statistically significant at a chosen level of significance and which has the minimum variance of residuals among the sub-models that have all terms significant at that level.

In general, neither version of stepwise regression yields the optimum model but in most cases the model obtained by either procedure comes very close to being optimum and in many cases is identical to that obtained by the costly method of enumerating all the solutions.

In those instances where one is interested in finding the optimum model, as defined above, the Gauss-Jordan algorithm greatly reduces the required computations. The optimum path of elimination for generating all possible stepwise combinations can be controlled by a "binary algorithm" described by Lotto [14], 1961, and Garside [6], 1965. The procedure is optimized so that the computations go through the fewest recursions. Despite this optimization, the computational labor is such that the procedure seems limited to handling fewer than twenty variables.

The paper by Efroymson contains mostly a description of the computational procedure. This report contains derivations of the pertinent mathematical equations related to the procedure including the recurrence formulas relating covariances of residuals, regression coefficients, and elements of the inverse of partitions of the covariance matrix. An improvement of the algorithm used by Efroymson is derived. This improved algorithm reduces the storage requirement by 50% thus allowing the analysis of larger models or the use of double precision arithmetic. This latter consideration is quite important when analysing models containing many variables. In addition, a numerical example is presented showing the differing results that can be obtained by the backward and forward versions of the procedure.

A computer program for BRLESC (Ballistic Research Laboratories Electronic Scientific Computer) which incorporates this procedure is described by the author and others in a previous report, BRL Report No. 1330, July 1966. The present report is an amplification of the statistical theory and computational procedures presented in that report in addition to the exposition of the improved algorithm.

II. MULTIPLE LINEAR REGRESSION

The theory of multiple linear regression and correlation is contained in the theory of "Linear Statistical Models" and can be found in many widely used texts such as that by Graybill [7]. The concept of a linear model is fundamental to the ensuing exposition and hence the definition found in Graybill is listed. By a linear model is meant "an equation that relates random variables, mathematical variables, and parameters and that is linear in the parameters and in the random variables." Linear models are classified into several categories depending on the distribution of the variables, the presence and nature of errors when observing the variables, and in the nature of the variables themselves, i.e., whether the variables are mathematical variables or random variables. The equation relating the variables is written in the form

$$X_n = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_{n-1} X_{n-1}. \quad (1)$$

The variables X_1, X_2, \dots, X_{n-1} are referred to as "independent variables" and X_n as the dependent variable. In some instances one is interested in polynomial or curvilinear models and the variables X_1, X_2, \dots, X_{n-1} are not necessarily independent in the probability sense. For example the model

$$X_2 = b_1 X_1 + b_2 \cos X_1 + b_3 e^{X_1} \quad (2)$$

is curvilinear, i.e., linear in the parameters b_1 , b_2 and b_3 even though nonlinear in X_1 . This model fits into the framework of Equation (1) when the transformations $X_2 = \cos X_1$ and $X_3 = e^{X_1}$ are introduced. This model is contrasted with the model

$$X_2 = b_1 e^{b_2 X_1} + b_3 \cos b_4 X_1 \quad (3)$$

which is nonlinear in the parameters b_1 , b_2 , b_3 and b_4 and cannot be linearized by transformations. This problem is one of nonlinear regression and is not discussed further in this report.

In multiple linear regression one is interested in obtaining an estimate of the b_1 which will yield a "prediction equation" represented by Equation (1) which best fits a set of observations. The m sets of observations of X_n , the dependent variable, and of X_1, X_2, \dots, X_{n-1} can be written as a matrix x_{ij} , $i = 1, 2, \dots, m$, $j = 1, 2, \dots, n$. When the variables are measured about their respective means, Equation (1) can be written

$$X_n - \bar{X}_n = b_1 (X_1 - \bar{X}_1) + b_2 (X_2 - \bar{X}_2) + \dots + b_{n-1} (X_{n-1} - \bar{X}_{n-1}). \quad (4)$$

The coefficient b_0 in Equation (1) is obtained from the relationship

$$b_0 = \bar{X}_n - \sum_{i=1}^{n-1} b_i \bar{X}_i. \quad (5)$$

Hereafter the variables will be assumed to be measured about their respective means and the quantity X_1 will be used to represent $X_1 - \bar{X}_1$.

For a particular observation Equation (4) takes the form

$$x_{jn} = b_1 x_{j1} + b_2 x_{j2} + \dots + b_{n-1} x_{j,n-1} + e_j. \quad (6)$$

e_j is a residual and is the difference between the predicted value and the observed value of X_n^* . The least-squares method of estimating the coefficients b_i is based on the minimization of the sum of the squares of the residuals, denoted as E^2 .

$$\begin{aligned} E^2 &= \sum_{j=1}^m e_j^2 \\ &= \sum_{j=1}^m (x_{jn} - b_1 x_{j1} - b_2 x_{j2} - \dots - b_{n-1} x_{j,n-1})^2 \end{aligned} \quad (7)$$

This minimization is achieved by taking partial derivatives of E^2 with respect to each of the b_k and equating each of these $(n-1)$ equations to zero. This leads to the normal equations

$$\begin{aligned} \sum_{j=1}^m x_{jk} (x_{jn} - b_1 x_{j1} - b_2 x_{j2} - \dots - b_{n-1} x_{j,n-1}) &= 0. \quad (8) \\ k &= 1, 2, \dots, n-1 \end{aligned}$$

The normal equations can be written in matrix form

$$X'X B = X'Y. \quad (9)$$

X is the $m \times (n-1)$ matrix of observations of the independent variables, X' its transpose, Y is the $m \times 1$ matrix of observations of the dependent

*It should be noted that the variables X_i , $i = 1, 2, \dots, n$, are assumed to be measured without error.

variable and B is the column vector of $(n-1)$ regression coefficients. The solution of the normal equations to obtain the regression coefficients is given as

$$B = \begin{pmatrix} b_1 \\ b_2 \\ . \\ . \\ b_{n-1} \end{pmatrix} = (X'X)^{-1} X'Y, \quad (10)$$

where $(X'X)^{-1}$ is the inverse of the matrix $X'X$. The normal equations can be solved by any of several algorithms for the solution of systems of linear equations, however, the Gauss-Jordan algorithm is used in stepwise multiple regression for reasons that will become apparent.

III. COMPUTATIONAL CONSIDERATIONS IN MULTIPLE LINEAR REGRESSION

The most severe computational problem occurring in multiple linear regression is the formation and solution of the normal equations. For any problem containing more than a few variables and observations this problem can become too laborious for desk calculation and the use of high speed computers is very desirable. As a consequence, generalized library programs for doing multiple regression computations are widely available and can be obtained in most computing facilities. In general it is desirable for these programs to do more than compute regression coefficients and variance of residuals, they should also provide associated statistical data that could be used for significance tests, computing prediction intervals, etc. These considerations are discussed by Slater [6], 1961 and by Healy [11], 1963. These programs should be designed as efficiently as possible to keep the computation time reasonably small. Since the Gauss-Jordan algorithm provides the solution to $(n-1)$ regression models en route to solving the complete problem at essentially no significant increase in cost compared to other algorithms, it seems wherever any library program for multiple regression is prepared, the program should incorporate the stepwise scheme. Such a program could then be used either to provide only the complete solution or to select the significant variables for inclusion in the output model.

The programming effort required to include the optional capabilities for both forward stepwise regression and backward stepwise regression is relatively small compared to the total programming effort required to prepare either program. For this reason it seems worthwhile that a well designed computer program should provide a capability for both types of computations. The relative advantages and disadvantages of the two procedures will be discussed in a later section. The effort required to prepare the matrix elements to begin the backward stepwise regression is identical to the effort required to perform a complete forward regression. Because of this it seems advisable that when the backward option is selected, the program should be controlled in a manner which yields the results of a normal forward regression as a by-product. When proceeding forward the various solutions obtained may correspond to models of the form:

$$\begin{aligned}
 X_n &= b_0 + b_1 X_1 \\
 X_n &= b'_0 + b'_1 X_1 + b'_3 X_3 \\
 X_n &= b''_0 + b''_1 X_1 + b''_3 X_3 + b''_7 X_7
 \end{aligned}
 \tag{11}$$

At each stage the program, at a minimum, should print the standard deviation of residuals and identify the variables entered or removed. This information can then prove to be invaluable if one chooses a simpler model than the one finally selected by the stepwise regression procedure.

IV. MATHEMATICAL BASIS OF THE STEPWISE REGRESSION

The mathematical basis of the stepwise regression is that the transformation rules of the Gauss-Jordan algorithm correspond to recurrence relations that exist between covariances of residuals, regression coefficients, and inverse elements of partitions of the covariance matrix. These relations can readily be derived by taking advantage of Yule's notation as described by Kendall [13]. In this notation the regression Equation (1) is written as follows:

$$X_n = b_{n1.23\dots n-1} X_1 + b_{n2.13\dots n-1} X_2 + \dots + b_{n,n-1.12\dots n-2} X_{n-1} \quad (12)$$

The first subscript of each b is that corresponding to the dependent variable, the second subscript corresponding to the variable attached to the regression coefficient. These two subscripts are called the primary subscripts. The remaining subscripts on the right of the period are those of the remaining variables and are called secondary subscripts. The entire collection of subscripts for those variables that are in regression is thus represented by those subscripts to the right of the period with the addition of the subscript to the immediate left of the period. It should be noted that on a regression coefficient neither of the primary subscripts can ever be included in the secondary subscripts.

In a similar notation the residuals are denoted as

$X_{n.12...(n-1)}$. The subscript to the left of the period is that of the dependent variable and those to the right are the subscripts of the independent variables in the regression. Since regressions containing fewer than the $(n-1)$ independent variables will be of interest it is necessary to introduce the following notation. The subscript q will be used to represent the collection of subscripts 1 through $(k-1)$ with the exclusion of i and j , i.e.,

$$q = 1, 2, \dots (i-1)(i+1) \dots (j-1)(j+1) \dots (k-1).$$

Any variable can be considered as the dependent variable, e.g., the residuals $X_{i.q}$ and $X_{j.q}$ will be utilized in deriving the recurrence relations. The covariance of the variables X_i and X_j is defined as

$$s_{ij} = \sum^* X_i X_j / f$$

where f is the degrees of freedom and the summation extends over the m data points. For the present f will be defined as m and therefore does not vary as the number of variables in regression varies. The covariance of residuals is defined as

$$s_{ij.q} = \sum X_{i.q} X_{j.q} / f$$

The secondary subscripts of a covariance indicate the variables in the regression. When using this notation neither of the primary subscripts

*Hereafter, unless denoted otherwise, all summations extend over the m data points.

can be included in the secondary subscripts. The collection of variables whose subscripts are contained in q , is always assumed to be in regression, however additional variables such as X_i and X_j (whose subscripts are not contained in q) may also be in regression. For a covariance the presence of this situation is denoted as follows:

$$s_{kk.qij} = \sum x_{k.12...(k-1)}^2$$

Similar notation will be used for the regression coefficients and for elements of the inverse of partitions of the covariance matrix.

In the above notation, the normal equations (for the entire collection of variables) can be written in the form

$$\sum x_{n.12...n-1} x_r = 0, r = 1, 2, \dots, n-1 \quad (13)$$

or equivalently

$$s_{1r} b_{n1.23...(n-1)} + s_{2r} b_{n2.13...(n-1)} + \dots + s_{(n-1)r} b_{n(n-1).12...(n-2)} = s_{nr}, r = 1, 2, \dots, (n-1). \quad (14)$$

The complete covariance matrix is:

$$S = \begin{pmatrix} s_{11} & s_{12} & \dots & s_{1n} \\ s_{21} & s_{22} & \dots & s_{2n} \\ \dots & \dots & \dots & \dots \\ s_{n1} & s_{n2} & \dots & s_{nn} \end{pmatrix} \quad (15)$$

This matrix corresponds to the augmented matrix of coefficients usually considered in solving a system of linear equations with the addition

of the nth row. The nth row is added so that the variance of residuals, $s_{nn.q}$ will be made available through the recurrence formulas, thus avoiding the need for computing residuals at each stage.

Derivation of Recurrence Formulas

In deriving the recurrence formulas it is convenient to take note of Kendall's [13] three observations:

(a) The covariance of any residual and any variable is zero provided that the subscript of the variable occurs among the secondary subscripts of the residual, i.e., $\sum X_i X_{j.qi} = 0$.

(b) The covariance of any two residuals is zero provided that the subscripts of either residual are contained in the secondary subscripts of the other, i.e., $\sum X_{i.q} X_{j.qi} = 0$.

(c) The covariance of any two residuals is unaltered by omitting any or all terms in either residual whose secondary subscripts are contained in the secondary subscripts of the other residual, i.e., $\sum X_{i.q} X_{j.qi} = \sum X_{i.q} (X_j - b_{ji.q} X_i)$.

Statement (a) is merely a statement of the normal equations. (b) and (c) arise as a consequence of (a).

The actual value of a recurrence formula in computation is dependent upon the availability of all the elements entering in the recurrence except the one to be determined. With this in mind the

ensuing recurrences are derived and their relationship to the Gauss-Jordan algorithm will be exhibited. Furthermore it will be shown that the algorithm can be used without modification in a backwards recursion, i.e., once a term is in regression it can be removed by the same algorithm. Altogether 18 recurrence relations are of interest. Nine of these correspond to the introduction of variables in regression and the remaining nine correspond to the removal of variables from the regression. It will be shown that these 18 recurrence formulas are equivalent to the four rules of the Gauss-Jordan algorithm. The elements of the derivations do not necessitate any particular sequencing of the digits in q (the sequence has been assumed for simplicity) and hold true for arbitrary i, j and k . The presence of X_i, X_j and X_k in regression (or not) will be denoted by the notation introduced previously.

From (c)

$$\sum X_{k.q} X_{j.qk} = 0 = \sum X_{k.q} (X_j - b_{jk.q} X_k).$$

Also $\sum X_{k.q} X_j = \sum X_{k.q} X_{j.q}$ and $\sum X_{k.q} X_k = \sum X_{k.q} X_{k.q}$.

Hence

$$\sum X_{k.q} X_{j.q} = b_{jk.q} \sum X_{k.q}^2.$$

Dividing by f

$$b_{jk.q} = s_{kj.q}/s_{kk.q} = s_{jk.q}/s_{kk.q}. \quad (16)$$

As shown later, it is useful to define a new quantity $d_{ik,q}$ as follows:

$$d_{ik,q} = -b_{ik,q} = -s_{ik,q}/s_{kk,q} \quad (17)$$

Again from (c)

$$\begin{aligned} \sum X_{i,qk} X_{j,qk} &= \sum X_{i,q} X_{j,qk} \\ &= \sum X_{i,q} (X_j - b_{jk,q} X_k) \\ &= \sum X_{i,q} X_{j,q} - b_{jk,q} \sum X_{i,q} X_{k,q} \end{aligned}$$

or equivalently

$$s_{ij,qk} = s_{ij,q} - b_{jk,q} s_{ik,q}$$

Substituting for $b_{jk,q}$ from Equation (16)

$$s_{ij,qk} = s_{ij,q} - s_{ik,q} s_{kj,q}/s_{kk,q} \quad (18)$$

From (16) and (17)

$$\begin{aligned} b_{ji,qk} &= s_{ij,qk}/s_{ii,qk} \\ &= \frac{s_{ij,q} s_{kk,q} - s_{ik,q} s_{kj,q}}{s_{ii,q} s_{kk,q} - s_{ik,q} s_{ki,q}} \\ &= b_{ji,q} - \frac{s_{ij,q}}{s_{ii,q}} + \frac{s_{ij,q} s_{kk,q} - s_{ik,q} s_{kj,q}}{s_{ii,q} s_{kk,q} - s_{ik,q} s_{ki,q}} \end{aligned}$$

$$= b_{ji.q} - \frac{s_{ik.q} s_{kj.q} s_{ii.q} - s_{ki.q} s_{ij.q}}{s_{ii.q} s_{ii.q} s_{kk.q} s_{ik.q} s_{kk.q}}$$

or

$$b_{ji.qk} = b_{ji.q} - b_{ki.q} s_{kj.qi} / s_{kk.qi} \quad (19)$$

Equivalently

$$-b_{ij.qk} = -b_{ij.q} - (-b_{kj.q}) s_{ik.qj} / s_{kk.qj}$$

Hence

$$d_{ij.qk} = d_{ij.q} - s_{ik.qj} d_{kj.q} / s_{kk.qj} \quad (20)$$

Elements of the Inverse Matrix

Consider the partition of the covariance matrix formed by taking all the rows and columns of indices q, i, j, k . Denote the determinant of this matrix as R and the cofactor of the element s_{ij} as R_{ij} . Since the covariance matrix is symmetrical, $R_{ij} = R_{ji}$. From Craemer's rule

$$b_{ij.qk} = -R_{ij} / R_{ii} \quad (21)$$

$$\begin{aligned} s_{ii.qjk} &= \sum x_{i.qjk}^2 / f = \sum x_{i.qjk} x_i / f \\ &= s_{ii} - \sum_{t=q,j,k} b_{it.12...(i-1)(i+1)...(t-1)(t+1)...k} s_{it} \\ &= s_{ii} + 1/R_{ii} \sum_{t=q,j,k} s_{it} R_{it} \\ &= \sum_{t=q,j,k} s_{it} R_{it} / R_{ii} \end{aligned}$$

From the Laplace expansion theorem

$$R = \sum_{t=q,i,j,k} s_{it} R_{it}.$$

Hence

$$s_{ii.qjk} = R/R_{ii}. \quad (22)$$

From Equation (16)

$$\begin{aligned} s_{ij.qk} &= b_{ji.qk} s_{ii.qk} \\ &= - (R_{ji}/R_{jj}) (R_{jj}/R_{ii-jj}). \end{aligned}$$

R_{hi-jk}^* is the cofactor of the second order minor in R which is obtained by striking out row h and column i and then row j and column k .

$$s_{ij.qk} = - R_{ji}/R_{ii-jj} = - R_{ij}/R_{ii-jj}. \quad (23)$$

The i,j th element of the inverse of the partition of the covariance matrix defined above is denoted as $c_{ij.qijk}$. The only inverse elements which will be of interest are those elements which are inverse elements of partitions defined by taking the rows and columns subscripted by the subscripts of the variables in regression. Hence the primary subscripts of the inverse elements will always be included in the secondary subscripts. As in the case of covariances, the secondary subscripts will denote the variables in regression. From fundamentals of matrix algebra

*This notation is taken from Gutman [8].

$$\begin{aligned}
c_{ik.qijk} &= R_{ki}/R = R_{ik}/R \\
&= (R_{ki}/R_{kk})(R_{kk}/R).
\end{aligned} \tag{24}$$

Hence

$$c_{ik.qijk} = -b_{ki.qj}/s_{kk.qij} \tag{25}$$

Similarly

$$\begin{aligned}
c_{kj.qijk} &= c_{jk.qijk} = -b_{kj.qi}/s_{kk.qij} \\
&= d_{kj.qi}/s_{kk.qij}
\end{aligned} \tag{26}$$

and

$$\begin{aligned}
c_{kk.qijk} &= R_{kk}/R = 1/(R/R_{kk}) \\
&= 1/s_{kk.qij}.
\end{aligned} \tag{27}$$

From Equation (25)

$$\begin{aligned}
c_{ij.qijk} &= -b_{ji.qk}/s_{jj.qik} \\
&= -\frac{b_{ji.q} s_{kk.qi} - b_{ki.q} s_{kj.qi}}{s_{jj.qi} s_{kk.qi} - s_{jk.qi} s_{kj.qi}} \\
&= c_{ij.qij} - \frac{b_{ji.q}}{s_{jj.qi}} - \frac{b_{ji.q} s_{kk.qi} - b_{ki.q} s_{kj.qi}}{s_{jj.qi} s_{kk.qi} - s_{jk.qi} s_{kj.qi}} \\
&= c_{ij.qij} + \frac{s_{kj.qi}/s_{jj.qi} (b_{ki.q} - b_{ji.q} s_{jk.qi}/s_{jj.qi})}{s_{kk.qi} - s_{jk.qi} s_{kj.qi}/s_{jj.qi}}
\end{aligned}$$

or

$$\begin{aligned}
c_{ij.qijk} &= c_{ij.qij} + b_{ki.qj} b_{kj.qi}/s_{kk.qij} \\
&= c_{ij.qij} - b_{ki.qj} d_{kj.qi}/s_{kk.qij}.
\end{aligned} \tag{28}$$

The formulas derived to this point are those for forward recursion, or for the addition of variables into the regression. Similar formulas are now derived for backward recursion.

From Equation (25)

$$b_{ki.qj} = - c_{ik.qijk} s_{kk.qij} = - c_{ik.qijk} / c_{kk.qijk} \quad (29)$$

Similarly

$$d_{kj.qi} = c_{jk.qijk} / c_{kk.qijk} \quad (30)$$

From Equation (28)

$$c_{ij.qij} = c_{ij.qijk} + b_{ki.qj} d_{kj.qi} / s_{kk.qij}$$

Substituting for

$$b_{ki.qj} = - c_{ik.qijk} s_{kk.qij} \quad \text{and}$$

$$d_{kj.qi} = c_{jk.qijk} / c_{kk.qijk}$$

$$c_{ij.qij} = c_{ij.qijk} - c_{ik.qijk} c_{jk.qijk} / c_{kk.qijk} \quad (31)$$

From Equation (18)

$$\begin{aligned} s_{ij.q} &= s_{ij.qk} + s_{ik.q} s_{kj.q} / s_{kk.q} \\ &= s_{ij.qk} + b_{ik.q} s_{kk.q} b_{jk.q} s_{kk.q} / s_{kk.q} \end{aligned}$$

$$\text{or} \quad s_{ij.q} = s_{ij.qk} - d_{ik.q} b_{jk.q} / c_{kk.qk} \quad (32)$$

From Equation (27)

$$s_{kk.qij} = 1 / c_{kk.qijk} \quad (33)$$

From Equation (19)

$$\begin{aligned}
 b_{ji.q} &= b_{ji.qk} + b_{ki.q} s_{kj.qi} / s_{kk.qi} \\
 &= b_{ji.qk} - c_{ik.qik} s_{kk.qi} b_{jk.qi} / c_{kk.qik} s_{kk.qi} \\
 \text{or } b_{ji.q} &= b_{ji.qk} - c_{ik.qik} b_{jk.qi} / c_{kk.qik}
 \end{aligned} \tag{34}$$

Similarly

$$\begin{aligned}
 -b_{ij.q} &= -b_{ij.qk} - c_{jk.qjk} (-b_{ik.qj}) / c_{kk.qjk} \\
 \text{or } d_{ij.q} &= d_{ij.qk} - d_{ik.qj} c_{jk.qjk} / c_{kk.qjk}
 \end{aligned} \tag{35}$$

From Equation (16)

$$s_{kj.q} = b_{jk.q} s_{kk.q} = b_{jk.q} / c_{kk.qk}.$$

Similarly

$$s_{ik.q} = b_{ik.q} / c_{kk.qk} = -d_{ik.q} / c_{kk.qk}. \tag{37}$$

The eighteen recurrence formulas are listed in a convenient order on the following page. The successive application of these formulas to appropriate matrix elements is the basis of stepwise multiple linear regression. The matrix elements are continually replaced at each stage by the matrix elements of the new stage. The initial matrix is the covariance matrix, equation (15). Each stage is characterized by the presence of a particular set of independent variables in the regression. In practice the variables will not enter the regression in sequence, but in an order determined by their ability to reduce the variance of residuals. For the present we can assume that as the

List of Recurrence Formulas

1. $c_{ij.qijk} = c_{ij.qij} - b_{ki.qj} d_{kj.qi} / s_{kk.qij}$
2. $c_{ik.qijk} = - b_{ki,qj} / s_{kk.qij}$
3. $b_{ji.qk} = b_{ji.q} - b_{ki.q} s_{kj.qi} / s_{kk.qi}$
4. $c_{kj.qijk} = d_{kj.qi} / s_{kk.qij}$
5. $c_{kk.qijk} = 1 / s_{kk.qij}$
6. $b_{jk.q} = s_{kj.q} / s_{kk.q}$
7. $d_{ij.qk} = d_{ij.q} - d_{kj.q} s_{ik.qj} / s_{kk.qj}$
8. $d_{ik.q} = - s_{ik.q} / s_{kk.q}$
9. $s_{ij.qk} = s_{ij.q} - s_{ik.q} s_{kj.q} / s_{kk.q}$
10. $c_{ij.qij} = c_{ij.qijk} - c_{ik.qijk} c_{jk.qijk} / c_{kk.qijk}$
11. $b_{ki.qj} = - c_{ki.qj} / c_{kk.qijk}$
12. $b_{ji.q} = b_{ji.qk} - c_{ik.qi} b_{jk.qi} / c_{kk.qik}$
13. $d_{kj.qi} = c_{kj.qijk} / c_{kk.qijk}$
14. $s_{kk.qij} = 1 / c_{kk.qijk}$
15. $s_{kj.q} = b_{jk.q} / c_{kk.qk}$
16. $d_{ij.q} = d_{ij.q} - d_{ik.qj} c_{jk.qjk} / c_{kk.qjk}$
17. $s_{ik.q} = - d_{ki.q} / c_{kk.qk}$
18. $s_{ij.q} = s_{ij.qk} - d_{ik.q} b_{jk.q} / c_{kk.qk}$

variables enter the regression they are reordered. The end effect (after the reordering) is that the variables are introduced into the regression in the order X_1, X_2, \dots, X_k , hence, the k 'th stage is characterized by the presence of X_1, X_2, \dots, X_k in regression.

Theorem on Stepwise Multiple Linear Regression

Consider the sequence of matrices A_0, A_1, \dots, A_{n-1} . A_0 is the covariance matrix, Equation (15). A_k ($k = 1, 2, \dots, n-1$) is the matrix formed by applying the transformation

$$\begin{aligned} a_{ij}^k &= a_{ij}^{k-1} - a_{ik}^{k-1} a_{ki}^{k-1} / a_{kk}^{k-1}, \quad i = 1, 2, \dots, (k-1)(k+1) \dots, n \\ &\quad j = 1, 2, \dots, (k-1)(k+1) \dots, n \\ a_{ik}^k &= -a_{ik}^{k-1} / a_{kk}^{k-1} \quad i = 1, 2, \dots, (k-1)(k+1) \dots, n \quad (38) \\ a_{kj}^k &= a_{kj}^{k-1} / a_{kk}^{k-1} \quad j = 1, 2, \dots, (k-1)(k+1) \dots, n \\ a_{kk}^k &= 1/a_{kk}^{k-1} \quad i = j = k \end{aligned}$$

to the matrix A_{k-1} . a_{ij}^k is the i, j th element of the matrix A_k . Denote this transformation as T_k . The results of applying this transformation are contained in the following theorem:

THEOREM:

The matrix A_k contains four partitions, the respective partitions having elements as follows:

$$\begin{aligned} a_{ij} &= c_{ij.12\dots k}, \quad i = 1, 2, \dots, k, j = 1, 2, \dots, k \\ a_{ij} &= b_{ji.12\dots i-1,i+1\dots k}, \quad i = 1, 2, \dots, k, j = k+1, k+2 \dots n \quad (39) \end{aligned}$$

$$a_{ij} = d_{ij.12\dots i-1,i+1\dots k}, \quad i = k+1, k+2, \dots n, \quad j = 1, 2, \dots k$$

$$a_{ij} = s_{ij.12\dots k}, \quad i = k+1, k+2, \dots n, \quad j = k+1, k+2, \dots n$$

The proof is by induction. Assume that the theorem holds for A_{k-1} , then show that it necessarily must hold for A_k and furthermore that it holds for $k = 1$. The matrix A_{k-1} can be partitioned as follows:

$$A_{k-1} = \begin{pmatrix} A_{k-1,1} & A_{k-1,2} & A_{k-1,3} \\ A_{k-1,4} & A_{k-1,5} & A_{k-1,6} \\ A_{k-1,7} & A_{k-1,8} & A_{k-1,9} \end{pmatrix} \quad (40)$$

$$= \begin{pmatrix} c_{11} & c_{12} & \dots & c_{1,k-1} & b_{k1} & b_{k+1,1} & \dots & b_{n1} \\ c_{21} & c_{22} & \dots & c_{2,k-1} & b_{k2} & b_{k+1,2} & \dots & b_{n2} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ c_{k-1,1} & c_{k-1,2} & \dots & c_{k-1,k-1} & b_{k,k-1} & b_{k+1,k-1} & \dots & b_{n,k-1} \\ d_{k1} & d_{k2} & \dots & d_{k,k-1} & s_{kk} & s_{k,k+1} & \dots & s_{kn} \\ d_{k+1,1} & d_{k+1,2} & \dots & d_{k+1,k-1} & s_{k+1,k} & s_{k+1,k+1} & \dots & s_{k+1,n} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ d_{n1} & d_{n2} & \dots & d_{n,k-1} & s_{nk} & s_{n,k+1} & \dots & s_{nn} \end{pmatrix}$$

The secondary subscripts of the matrix have been omitted in A_{k-1} for brevity. The variables having subscripts 1, 2, ... k-1 are assumed to be in regression (due to the assumption that the theorem holds for A_{k-1}) and hence the appropriate secondary subscripts should be assumed to be attached to the various elements.

By inspection of the transformation T_k in relation to the elements stored in the nine partitions on which the transformation acts, it is seen that the application of T_k is identical to the application of the nine recurrence formulas 1 through 9. Furthermore, the application of the nine recurrence formulas to A_{k-1} is equivalent to replacing A_{k-1} with A_k . The same holds true for $k = 1$ and hence the proof is complete.

In a similar fashion it can be shown that as a consequence of the nine recurrence formulas for backwards recursion, i.e., 10 through 18, the application of T_k to A_k generates the matrix A_{k-1} .

The consequence of the above theorem can be generalized as follows: The collection of variables whose subscripts are represented by the values taken by k in the successive application of T_k are said to be in regression if k appears an odd number of times in the collection. Alternatively, a variable is said not to be in regression if its subscript does not appear in the collection, or if it appears an even number of times. The content of the matrix at any stage is as follows:

$a_{ij} = s_{ij}$.- when neither X_i nor X_j are in regression.

$a_{ij} = b_{ji}$.- when X_i is in regression but not X_j .

$a_{ij} = d_{ij}$.- when X_j is in regression but not X_i .

$a_{ij} = c_{ij}$.- when both X_i and X_j are in regression.

The secondary subscripts are those appropriate to the particular variables in the regression at that stage. A bookkeeping method for determining which variables are in regression will be described in Section VI.

The Correlation Matrix

For computational reasons it is desirable to transform the initial matrix A_0 (the covariance matrix) by dividing each element a_{ij} by $s_i s_j$ where $s_i = \sqrt{s_{ii}}$. The resulting matrix is a matrix of simple correlation coefficients r_{ij} , $i, j = 1, 2, \dots, n$ where

$$r_{ij} = a_{ij} / s_i s_j.$$

The diagonal elements of A_0 are then unity and the remaining elements are of a more uniform order of magnitude. The recurrence formulas remain valid as shown below:

Consider the regression equation

$$X_n / s_n = B_1 (X_1 / s_1) + B_2 (X_2 / s_2) + \dots + B_k (X_k / s_k).$$

By inspection it is seen that the covariance matrix for this system is equal to the correlation matrix defined above. The coefficients B_i are those that arise when A_0 is the correlation matrix. Hence the coefficient $b_{ni.q}$ is computed from the formula,

$$b_{ni.q} = B_{ni.q} s_n / s_i.$$

If $S_{ij.q}$ is a covariance arising from the transformed system, $s_{ij.q}$ can be recovered by the formula

$$s_{ij.q} = s_i s_j S_{ij.q}.$$

In particular, the variance of residuals is given by

$$s_{nn.q} = s_n^2 S_{nn.q}.$$

If $C_{ij.qij}$ is an inverse element of the transformed system then

$$c_{ij.qij} = C_{ij.qij} / s_i s_j.$$

V. SELECTING THE KEY VARIABLE

In forward stepwise regression the variable which is entered into regression is the one which yields the greatest reduction in the variance of residuals at that stage. For an arbitrary variable X_i that is not in regression it is seen from the recurrence formula 9 that the variance reduction is given by the quantity

$$V_i = a_{in} a_{ni} / a_{ii} = s_{in.q} s_{ni.q} / s_{ii.q}. \quad (41)$$

For an arbitrary variable X_i that is in regression the variance increase resulting from the removal of X_i from regression is given by 18.

$$-V_i = -a_{in} a_{ni} / a_{ii} = -d_{ni.q} b_{ni.q} / c_{ii.qi}. \quad (42)$$

For X_i not in regression V_i is positive and for X_i in regression V_i is negative.

After determining the key element it is necessary to test whether the variance reduction due to entering the key variable is statistically significant. By inspection of 9 it is seen that for $i = j = n$

$$s_{nn.qk} = s_{nn.q} (1 - s_{nk.q} s_{kn.q} / s_{nn.q} s_{kk.q}). \quad (43)$$

The quantity $(s_{nk.q} s_{kr.q} / r_{nn.q} s_{kk.q})^{1/2}$ is defined as the product moment coefficient of correlation between $X_{n.q}$ and $X_{k.q}$. This quantity is denoted as $r_{nk.q}$ and is often referred to as a partial correlation coefficient. Equation (43) can be written in the form

$$r_{nk.q}^2 = s_{nk.q} s_{kn.q} / s_{nn.q} s_{kk.q} = (s_{nn.q} - s_{nn.qk}) / s_{nn.q}. \quad (44)$$

By inspection $r_{nk.q}^2$ gives the fractional variance reduction obtained by adding X_k into the regression. If $r_{nk.q}$ is statistically different from zero, then we observe that the fractional variance reduction due to X_k is significant and that X_k should be brought into regression. For forward recursion $r_{nk.q}^2$ can be computed directly from the first expression of (44). For backwards recursion, i.e., to test whether a variable X_k can be removed from regression, $r_{nk.q}^2$ can be computed from the formula

$$r_{nk.q}^2 = V_k / (s_{nn.qk} + V_k). \quad (45)$$

A test of significance for $r_{nk.q}$ is listed by Graybill [7]. If the true coefficient $\bar{r}_{nk.q}$, for which $r_{nk.q}$ is an estimate, is zero the quantity

$$t = r_{nk.q} (f-2)^{1/2} / (1 - r_{nk.q}^2)^{1/2} \quad (46)$$

is distributed as the Student t distribution. A test of the hypothesis $r_{nk.q} \neq 0$ against the alternative $r_{nk.q} = 0$ is performed as follows: The quantity t is compared against the one-tailed t statistic, $t(f-2, c)$ appropriate to the degrees of freedom, f , and the confidence level, c .

The hypothesis is accepted if $t > t(f-2, c)$.

The test is used in two ways:

(A) At the beginning of a stage V_i is computed for all subscripts, $i = 1, 2, \dots, n-1$. The largest positive V_i identifies the key variable which should be tested for entering into the regression. The quantity $r_{nk.q}$ is computed using Equation (44) and the t test described above is performed. If $t > t(f-2, c)$ the variable X_k is entered into regression by performing the transformation T_k .

(B) The second part of the stage begins by again computing V_i for all i . The negative V_i identify the variables that are not in regression. The negative V_i of smallest magnitude identifies the key variable to test for removal. $r_{nk.q}$ is computed using Equation (45). If $t > t(f-2, c)$ the correlation is significant and the variable X_k should remain in regression. If $t < t(f-2, c)$ the variable can be removed from regression without significantly increasing the variance of residuals. X_k is removed from the regression by applying T_k . The procedure is repeated until all insignificant variables have been removed.

The modification of (A) and (B) above for backward regression is quite simple. Initially the recursion is controlled to proceed all the way forward, yielding the inverse of the covariance matrix. On the way back, after any variable is removed, the determination is made as to whether a variable removed previously has become significant, if so it is reentered. If not, then the least significant variable in

regression is removed, provided again that the resulting variance increase is not significant. As in the forward version, the procedure continues until the equilibrium point is reached.

VI. IMPROVEMENT OF THE ALGORITHM

The algorithm described by Efroymson requires n^2 words of storage for the covariance matrix and the successive matrices that are generated as the regression proceeds. For problems requiring only a few variables in the candidate model, this storage requirement creates no difficulty on modern computing machinery. The author has been involved in problems (see for example BRL Report No. 1348, [2]) where it was necessary to examine candidate models containing 96 variables. Fortunately the machine used on this problem, the Ballistic Research Laboratories BRLESC has over 30,000 words of built-in double precision storage, i.e., the standard word length in this computer is 68 binary bits or approximately 20 decimal digits. Most commercial machines have word lengths of only 8 or 10 decimal digits. The experience of various computing facilities on large scale matrix problems done on commercial machines is that double precision computations are required to avoid the computational problem associated with roundoff. The details of this roundoff phenomena associated with polynomial models is discussed by Ralston [15], page 233.

The necessity of doing a stepwise multiple regression program in double precision reduces the available storage by a factor of two and accordingly limits the size of the model which can be analyzed by

a factor of the square root of two. The modified algorithm derived below has been implemented in the BRLESC program described in [3] and requires only $(n^2 + 7n - 2)/2$ words of storage. In addition the computations related to the application of the recursion formulas is halved thus requiring less computer time.

In problems involving symmetric matrices it is common to take advantage of the symmetry to reduce computations and storage. This is especially true of least-squares computations since the covariance matrix is symmetric. The matrices involved in stepwise multiple regression are not symmetric, but might be termed pseudo symmetric, i.e., $|a_{ij}| = |a_{ji}|$, the elements are symmetric in absolute value. Except for signs, all the statistical information stored in the matrix A_k is contained in the upper triangular part of the matrix and the diagonal. The justification for storing the lower triangular matrix (and subsequently operating on it) seemingly is that the signs contained in the lower triangular matrix are used to indicate which variables are in regression and which are not. To keep track of which variables are in regression one can store a sequence of numbers z_1, z_2, \dots, z_n . The presence of a variable X_i in regression is denoted by the presence of -1 in z_i . Initially z_1, z_2, \dots, z_n are all $+1$ to denote no variables in regression. As a variable X_i is entered into regression or removed z_i is multiplied by -1 . If z_i is operated on an even number of times this means that X_i was removed from regression as often as it was entered and hence is not in. This would be so indicated by z_i since z_i would be equal to $(-1)^{2r} = +1$. Alternatively if z_i is

operated on an odd number of times z_i is equal to $(-1)^{2r+1} = -1$.

This indicates X_i is in regression.

One additional problem remains. The transformation of elements in the upper triangular matrix using T_k involves elements which by storage implications are in the lower triangular matrix. Since it is desired to modify the algorithm so that the lower triangular matrix will not be stored, some method is needed to determine the signs of the elements below the diagonal. The elements $c_{ij} = c_{ji}$ and $s_{ij} = s_{ji}$. If a_{ij} is a regression coefficient $a_{ij} = b_{ji} = -d_{ij}$. Hence we note that $a_{ij} = -a_{ji}$ if either X_i or X_j are in regression, but $a_{ij} = a_{ji}$ if both are in regression or if neither are in regression. By inspection of T_k it is seen that the only elements involved in transforming a_{ij} are a_{ij} itself and other elements which lie either in row k or column k . This leads one to look for a way of "filling in" row k and column k below the diagonal with proper signs at the beginning of the stage. This is most conveniently done by storing the row and column in separate storage as elements t_{ij} . If a_{ij} is on or above the diagonal then $t_{ij} = a_{ij}$. Hence two rules are immediately apparent.

$$\begin{array}{lll} t_{kj} = a_{kj} & j = k, k+1, \dots, n & \text{Upper triangle row } k \\ t_{ik} = a_{ik} & i = 1, 2, \dots, k-1 & \text{Upper triangle column } k \end{array}$$

By inspection it is seen that t_{ij} is obtained in magnitude by a_{ji} and in sign by $z_i z_j$. This leads to the additional two rules

$$t_{kj} = z_k z_j a_{jk} \quad j = 1, 2, \dots, k-1, \quad \text{Lower triangle row } k$$

$$t_{ik} = z_i z_k a_{ki} \quad i = k+1, k+2, \dots, n. \quad \text{Lower triangle column } k$$

Equations (38) are then used to generate the new upper triangular matrix. The complete algorithm is as follows:

$$t_{kj} = a_{kj} \quad j = k, k+1, \dots, n$$

$$t_{ik} = a_{ik} \quad i = 1, 2, \dots, k-1$$

$$t_{kj} = z_k z_j a_{jk} \quad j = 1, 2, \dots, k-1$$

$$t_{ik} = z_i z_k a_{ki} \quad i = k+1, k+2, \dots, n$$

$$a'_{ij} = a_{ij} - t_{ik} t_{kj} / t_{kk} \quad i = 1, 2, \dots, k-1, k+1, \dots, n$$

$$j = i, i+1, \dots, k-1, k+1, \dots, n$$

$$a'_{kj} = t_{kj} / t_{kk} \quad j = k+1, k+2, \dots, n$$

$$a'_{ik} = -t_{ik} / t_{kk} \quad i = 1, 2, \dots, k-1$$

$$a'_{kk} = 1 / t_{kk} \quad i = j = k$$

$$z'_k = -z_k$$

The primes denote the elements of the new matrix.

VII. A COMPARISON OF FORWARD AND BACKWARD STEPWISE REGRESSION

Hamaker [10], 1962, compared forward and backward stepwise regression on data taken from Hald [9]. This data concerned the heat evolved during the hardening of cement. The problem involved four independent variables X_1 , X_2 , X_3 and X_4 . The optimum model in this problem contains the variables X_1 and X_2 . In Hamaker's version of "forward selection" the variables were entered into the regression in the order X_4 , X_1 , X_2 , X_3 and in his "backward elimination" the variables are eliminated in the order X_3 , X_4 , X_1 , X_2 . He concludes that if a model containing two variables were selected the forward version would yield the model containing X_4 and X_1 while the backward version would yield the optimum model containing the variables X_1 and X_2 . Hamaker made no provision for removing variables as they became insignificant and in fact, a forward procedure which does provide this capability would in this example have arrived at the optimum model. The author analysed Hald's data using the computer program described in [3] and obtained the results listed on the next page.

| STAGE | ACTION TAKEN | VARIABLES IN REGRESSION AT END OF STAGE | STD. DEV. OF RESIDUALS |
|-------|--------------|---|---------------------------|
| 0 | - | - | 15.04 |
| 1 | Add X_4 | X_4 | 8.96 |
| 2 | Add X_1 | X_4, X_1 | 2.73 |
| 3 | Add X_2 | X_4, X_1, X_2 | 2.31 |
| 4 | Remove X_4 | X_1, X_2 | 2.41 |

The decision to add or remove variables were made at the 95% level of significance. It is quite possible that at other levels of significance different results might be obtained and in fact in Section IX. an example is listed showing that even for a "perfect fit" model the forward version does not obtain the optimum model whereas the backward version does.

Abt* et al [1] discuss the forward and backward versions and attribute the occurrence of different results to the presence of "compounds". They define a compound as

a set of $\bar{N} \leq N$ independent variables plus the dependent variable when the error variance associated with all \bar{N} independent variables is smaller, by orders of magnitude, than the error variance associated with any subset of $\bar{N}-1$ independent variables.

Their discussion, however, seems to be based on a stepwise procedure which does not allow for the removal of terms in the forward version,

* Also discussed in a paper titled "On the Identification of the Significant Independent Variables in Linear Models" by Klaus Abt, soon to be published in *Metrika*. Dr. Abt provided the author a preprint of this paper.

nor for the subsequent addition of variables that have been eliminated in the backward version. The end result of a regression run on Abt et al's program as in Hamaker's example is an ordering of the variables in either a forward or backward ranking. The ranking in the end has really no meaning in regards to the relative importance of the variables' contributions to the variance reduction. The author, for example, has observed the following phenomenon: In six stages of a forward run, five stages consisted of removing variables that had entered earlier. In this problem, variables that in the end were insignificant would have been highly ranked had they not been tested for removal.

The objective in multiple linear regression analysis is the obtaining of a "prediction model" as near optimum as is practical, and the ordering as discussed above is of interest only in relation to the information it provides in achieving this end. In this context a provision for removing terms in the forward version seems to be more effective toward achieving this goal than a forward procedure which merely orders the variables in the sequence which produces the greatest reduction in the sum of squares of residuals. Similarly, the backward version should seemingly include a provision for reentering variables if they subsequently become significant after their removal.

The cost of running regression problems on today's modern machinery is so small that it seems for many problems one might fruitfully apply both versions for comparison. When many observations

are involved in relation to the number of variables the formation of the covariance matrix seems to comprise the bulk of the computation time. On a problem involving 96 variables and 1439 observations the BRLESC program [3] ran 5.34 minutes in the forward version, entering 21 variables before reaching equilibrium. When the program was modified to take advantage of the modified algorithm derived earlier this same problem ran in 4.90 minutes. From these figures it is estimated that the formation of the covariance matrix required about 4.5 minutes and that a complete forward regression would take approximately 2.0 minutes with a similar estimate for the time required to do a backward regression. Most problems are of a much smaller scale and running time considerations are usually unimportant.

ACKNOWLEDGMENT

The author acknowledges valuable criticism of this report by his thesis advisors, Dr. William Davis and Dr. H. B. Tingey of the University of Delaware.

REFERENCES

1. K. Abt, G. Gemmill, T. Herring, and R. Shade, DA-MRCA: A Fortran IV Program for Multiple Linear Regression, Technical Report No. 2035, U.S. Naval Weapons Laboratory, Dahlgren, Virginia, March 1966.
2. H. J. Breaux, The Computation of Firing Tables for Guided Missiles, Ballistic Research Laboratories Report No. 1348, November 1966.
3. H. J. Breaux, L. W. Campbell, and J. C. Torrey, Stepwise Multiple Regression-Statistical Theory and Computer Program Description, Ballistic Research Laboratories Report No. 1330, July 1966.
4. W. G. Cochran, The Omission or Addition of an Independent Variate In Multiple Linear Regression, Journal of the Royal Statistical Society, Suppl., 2, 1938.
5. M. A. Efronson, Multiple Regression Analysis, Mathematical Methods for Digital Computers, Edited by Ralston and Wilf, John Wiley and Sons, Inc., 1960.
6. M. J. Garside, The Best Sub-Set in Multiple Regression Analysis, Applied Statistics, Journal of the Royal Statistical Society, Vol. XIV, 1965.
7. F. A. Graybill, An Introduction to Linear Statistical Models, Vol. I, McGraw-Hill Book Company, Inc., 1961.
8. L. Gutman, A Note on the Derivation of Formulae for Multiple and Partial Correlation, Annals of Mathematical Statistics, Vol. IX, 1938.
9. A. Hald, Statistical Theory with Engineering Applications, John Wiley and Sons, Inc., New York, 1952.
10. H. C. Hamaker, On Multiple Regression Analysis, Neerlandica, 16, 31-56, 1962.
11. M. J. R. Healy, Programming Multiple Regression, Computer Journal, Vol. VI, 1963, 64.
12. P. Horst, Item Analysis by the Method of Successive Residuals, Journal of Experimental Education, Vol 2, 1934.
13. M. G. Kendall, Advanced Theory of Statistics, Vol. I, Charles Griffin and Company, London, 1943.

REFERENCES (Continued)

14. G. Lotto, On the Generation of all Possible Stepwise Combinations, Mathematics of Computation, Vol. 16, 1962.
15. A. Ralston, A First Course in Numerical Analysis, McGraw-Hill Book Company, 1965.
16. L. J. Slater, Regression Analysis, Computer Journal, Vol. IV, 1961, 62, Published Quarterly by the British Computer Society.

APPENDIX

Numerical Example*

The following example illustrates the point made earlier, that even for a "perfect fit" model the forward version of stepwise regression might not identify the optimum model. The linear model from which the data was generated is of the form

$$X_4 = 4X_1 - X_2 + 3X_3. \quad (49)$$

The matrix of observations is:

$$\begin{array}{cccc} X_1 & X_2 & X_3 & X_4 \\ \left(\begin{array}{cccc} 1 & 0 & 0 & 4 \\ 0 & 2 & -1 & -5 \\ -1 & 3 & 2 & -1 \\ 4 & 10 & 1 & 9 \\ 2 & 0 & 8 & 32 \end{array} \right) \\ \bar{X}_1 = 6/5 & \bar{X}_2 = 3 & \bar{X}_3 = 2 & \bar{X}_4 = 39/5 \end{array}$$

Rather than the covariance matrix, S , we begin with the matrix fS , denoted A_0 .

*This example was discovered by Mr. L. W. Campbell of the Ballistic Research Laboratories, Aberdeen Proving Ground, Maryland.

$$A_0 = 1/25 \begin{pmatrix} 370 & 475 & 150 & 1455 \\ 475 & 1700 & -400 & -1000 \\ 150 & -400 & 1250 & 4750 \\ 1455 & -1000 & 4750 & 21070 \end{pmatrix}$$

At the first stage the test quantities for the reduction in the sum of squares of residuals is given by

$$V_1 = a_{14} a_{41}/a_{11} = 1/25 (1455)^2/370 = 28.9,$$

$$V_2 = a_{24} a_{42}/a_{22} = 1/25 (1000)^2/1700 = 23.5,$$

$$V_3 = a_{34} a_{43}/a_{33} = 1/25 (4750)^2/1250 = 722.0.$$

Since V_3 is the largest of the three test quantities, X_3 becomes the key variable. To test whether this variable will significantly reduce the sum of squares of residuals we obtain the coefficient r_{43} .

$$r_{43}^2 = a_{43} a_{34}/a_{33} a_{44} = (4750)^2/(1250)(21070) = .857$$

$$t = \frac{r_{43}^2(f-2)}{1 - r_{43}^2}$$

$$= \frac{.857(3)}{.143} = 4.24$$

$$t(f-2, .95) = t(3, .95) = 2.35$$

Since $t > t(f-2, .95)$ the test for adding the variable indicates that X_3 (at the 95% level of confidence) should be brought into the regression. After operating on A_0 with the Gauss-Jordan algorithm with a_{33} as the pivot we obtain

$$A_1 = 1/25 \begin{vmatrix} 352 & 523 & -3 & 885 \\ 523 & 1572 & 8 & 520 \\ 3 & -8 & 1/2 & 95 \\ 885 & 520 & -95 & 3020 \end{vmatrix}$$

The test quantities are

$$v_1 = 1/25(885)^2/342 = 91.7,$$

$$v_2 = 1/25(520)^2/1572 = 68.7.$$

The key variable by inspection is X_1 .

$$r_{41.3}^2 = (885)/(342)(3020) = .758$$

$$t = \frac{.758(2)}{.342}^{1/2} = 2.10$$

$$t(f-2, .95) = t(2, .95) = 2.92$$

Since $t < t(f-2, .95)$ the test for addition fails and the variable X_1 is not entered into regression. This then is the equilibrium point and the model which a forward stepwise procedure would yield is

$$\bar{X}_4 - \bar{X}_4 = b_3 (X_3 - \bar{X}_3),$$

$$b_3 = a_{43} = 95/25 = .38,$$

$$b_0 = \bar{X}_4 - b_3 \bar{X}_3 = 39/5 - (2)95/25 = .2,$$

$$X_4 = .2 + .38 X_3.$$

Note that in this example no tests for removal were necessary.

It is not necessary to do the complete computations to exhibit the result for the backward version. One of the three variables,

(assume X_2) will be the key variable to test for removal. The partial correlation coefficient is computed from Equation (45).

$$r_{42.13}^2 = v_2 / (s_{44.123} + v_2)$$

Since $s_{44.123} = 0$, the coefficient is 1.0 indicating perfect correlation. This would be true for any of the three variables. Obviously, no variable is removed and the equilibrium point is established with all three variables in regression.

Recent Work in Europe

After the completion of this manuscript the author attended a seminar titled, "A New Computer Approach in Determining Optimum Regression in Multivariate Analysis." The lecturer was Dr. M. G. Kendall, the noted British statistician. The new approach referred to in the seminar title was a modification of the technique described by Lotto and Garside in enumerating the $2^N - 1$ regressions. Kendall and his coworkers have developed an algorithm which is more economical than the recursive generation of the $2^N - 1$ regressions by noting that it is possible to identify (without performing the computations) certain useless combinations which are demonstrably worse than combinations for which regressions have already been obtained. The details of this algorithm can be found in the paper "The Discarding of Variables in Multivariate Analysis" by E. M. L. Beale, M. G. Kendall and D. W. Mann, copies of which were distributed at the seminar*. This technique has

*This seminar was held on April 11, 1967 and sponsored by C-E-I-R Inc., 5272 River Road, Washington, D.C.

been called "partial enumeration" and its attractiveness in comparison to forward and backward stepwise regression was noted. It was pointed out, as was done earlier in this thesis, that stepwise regression does not in general lead to the optimum model. In this connection, reference was made to a paper by Oosterhoff* (1963) which contains an example for which the forward and backward methods lead to the same model, neither of which is optimum.

*Oosterhoff, J. (1963), On the Selection of Independent Variables in a Regression Equation, Report S 319 (VP23) Mathematisch Centrum, Amsterdam.

Unclassified

Security Classification

| DOCUMENT CONTROL DATA - R&D | | |
|---|----------------------|--|
| (Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified) | | |
| 1 ORIGINATING ACTIVITY (Corporate author) | | 2a REPORT SECURITY CLASSIFICATION |
| U.S. Army Ballistic Research Laboratories Aberdeen Proving Ground, Maryland | | Unclassified |
| | | 2b GROUP |
| 3 REPORT TITLE | | |
| ON STEPWISE MULTIPLE LINEAR REGRESSION | | |
| 4 DESCRIPTIVE NOTES (Type of report and inclusive dates) | | |
| 5 AUTHOR(S) (Last name, first name, initial) | | |
| Harold J. Breaux | | |
| 6 REPORT DATE | 7a TOTAL NO OF PAGES | 7b NO OF REFS |
| August 1967 | 58 | 16 |
| 8a CONTRACT OR GRANT NO. | | 9a ORIGINATOR'S REPORT NUMBER(S) |
| a PROJECT NO. RDT&E 1P014501A14B | | Report No. 1369 |
| c | | 9b OTHER REPORT NO(S) (Any other numbers that may be assigned this report) |
| d | | |
| 10 AVAILABILITY/LIMITATION NOTICES | | |
| Distribution of this document is unlimited. | | |
| 11 SUPPLEMENTARY NOTES This report is based on a master's thesis presented to the University of Delaware, Dept. of Statistics & Computer Science, June 67 | | 12 SPONSORING MILITARY ACTIVITY |
| | | U.S. Army Materiel Command Washington, D.C. |
| 13 ABSTRACT | | |
| <p>Stepwise multiple linear regression has proved to be an extremely useful computational technique in data analysis problems. This procedure has been implemented in numerous computer programs and overcomes the acute problem that often exists with the classical computational methods of multiple linear regression. This problem manifests itself through the excessive computation time involved in obtaining solutions to the 2^N-1 sets of normal equations that arise when seeking an optimum linear combination of variables from the subsets of the N variables. The procedure takes advantage of recurrence relations existing between covariances of residuals, regression coefficients, and inverse elements of partitions of the covariance matrix. The application of these recurrence formulas is equivalent to the introduction or deletion of a variable into a linear approximating function which is being sought as the solution to a data analysis problem. This report contains derivations of the recurrence formulas, shows how they are implemented in a computer program and includes an improved algorithm which halves the storage requirements of previous algorithms. A computer program for the BRLESC computer which incorporates this procedure is described by the author and others in a previous report, BRL Report No. 1330, July 1966. The present report is an amplification of the statistical theory and computational procedures presented in that report in addition to the exposition of the improved algorithm.</p> | | |

DD FORM 1 JAN 64 1473

Unclassified

Security Classification

| KEY WORDS | LINK A | | LINK B | | LINK C | |
|---|--------|----|--------|----|--------|----|
| | ROLE | WT | ROLE | WT | ROLE | WT |
| Multiple Regression Statistical Recurrence Formulas Correlation Linear Statistical Models Statistical Computer Program Curve Fitting | | | | | | |

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization (*corporate author*) issuing the report.

2a. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year, or month, year. If more than one date appears on the report, use date of publication.

7a. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8a. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers (*either by the originator or by the sponsor*), also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

- (1) "Qualified requesters may obtain copies of this report from DDC."
- (2) "Foreign announcement and dissemination of this report by DDC is not authorized."
- (3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through _____."
- (4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through _____."
- (5) "All distribution of this report is controlled. Qualified DDC users shall request through _____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring (paying for) the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as (TS), (S), (C), or (U).

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.